



①⑨ BUNDESREPUBLIK

DEUTSCHLAND



DEUTSCHES

PATENTAMT

⑫ **Offenlegungsschrift**
⑩ **DE 195 16 776 A 1**

⑳ Aktenzeichen: 195 16 776.7
㉑ Anmeldetag: 10. 5. 95
㉒ Offenlegungstag: 14. 11. 96

㉓ Int. Cl.⁶:
C 12 N 15/11
C 12 N 1/00
C 12 N 5/10
C 07 K 14/435
C 07 K 16/18
A 61 K 48/00
A 61 K 39/395
A 01 K 67/027
G 01 N 33/50

DE 195 16 776 A 1

㉔ Anmelder:

Boehringer Ingelheim International GmbH, 55218
Ingelheim, DE

㉕ Erfinder:

Jenuwein, Thomas, Dr., Wien, AT; Laible, Götz, Dr.,
Wien, AT

㉖ Chromatin-Regulatorgene

㉗ Die Deregulation von Chromatinregulator-Genen, die eine SET-Domäne aufweisen, spielt eine Rolle bei bestimmten Krebserkrankungen. Diese Gene können somit, insbesondere die SET-Domäne als solche, als Grundlage für die Diagnose und Therapie dieser Erkrankungen dienen.

DE 195 16 776 A 1

Die vorliegende Erfindung bezieht sich auf Gene, die eine Rolle bei der Regulation von Chromatin spielen, und ihre Verwendung für die Therapie und Diagnostik.

Die funktionelle Organisation von eukaryotischen Chromosomen in Centromere, Telomere sowie in eu- und heterochromatische Regionen stellt einen entscheidenden Mechanismus zur Gewährleistung der genauen Replikation und Verteilung der genetischen Information bei jeder Zellteilung dar. Im Gegensatz dazu sind Tumorzellen häufig durch chromosomale Rearrangements, Translokationen und Aneuploidie charakterisiert (Solomon et al., 1991; Pardue, 1991). Obwohl die Mechanismen, die zu einer erhöhten Chromosomeninstabilität in Tumorzellen führen, noch nicht geklärt sind, haben es in jüngster Zeit eine Reihe von experimentellen Systemen, beginnend mit telomerischen Positionseffekten in Hefe (Renauld et al., 1993; Buck und Shore, 1995; Allshire et al., 1994), über Positionseffekt-Variation (PEV) in *Drosophila* (Reuter und Spierer, 1992), bis zur Analyse von Translokationsbruchpunkten in humanen Leukämien (Solomon et al., 1991; Cleary, 1991), ermöglicht, einige chromosomale Proteine zu identifizieren, die an der deregulierten Proliferation ursächlich beteiligt sind.

Erstens wurde festgestellt, daß die Überexpression einer verkürzten Version des SIR4-Proteins zu einer verlängerten Lebensdauer in Hefe führt (Kennedy et al., 1995). Da SIR-Proteine zum Entstehen multimerer Komplexe an den stillen "Mating Type Loci" und am Telomer beitragen, könnte es sein, daß überexprimiertes SIR4 mit diesen heterochromatinartigen Komplexen interferiert, was schließlich zu einer unkontrollierten Proliferation führt. Diese Annahme steht im Einklang mit der Häufigkeit des Auftretens einer deregulierten Telomerenlänge in den meisten humanen Krebsarten (Counter et al., 1992).

Zweitens wurden mittels genetischer Analysen von PEV in *Drosophila* eine Reihe von Genprodukten identifiziert, die die Chromatinstruktur an heterochromatischen Positionen und innerhalb des homeotischen Genclusters verändern (Reuter und Spierer, 1992). Mutationen einiger dieser Gene, wie *modulo* (Garzino et al., 1992) und *polyhomeotic* (Smouse und Perrimon, 1990), können in *Drosophila* die deregulierte Zellproliferation oder den Zelltod verursachen. Drittens wurden Säugetierhomologe von sowohl Aktivatoren (*trithorax* oder *trx*-Gruppe) als auch Repressoren (z. B. *polycomb* oder *Pc*-Gruppe) der Chromatinstruktur von homeotischen *Drosophila*-Selektorgenen beschrieben. Unter diesen hat sich vom humanen HRX/ALL-1 (*trx*-Gruppe) gezeigt, daß es an der durch Translokation induzierten Leukämogenese beteiligt ist (Tkachuk et al., 1992; Gu et al., 1992), und daß die Überexpression des Maus-*bmi* (*Pc*-Gruppe) zum Entstehen von Lymphomen führt (Haupt et al., 1991; Brunk et al., 1991; Alkema et al., 1995). Ein Modell für die Funktion chromosomaler Proteine läßt darauf schließen, daß diese multimeren Komplexe bilden, welche in Abhängigkeit von der Ausgewogenheit zwischen Aktivatoren und Repressoren im Komplex den Kondensationsgrad der umliegenden Chromatinregion bestimmen (Locke et al., 1988). Eine Verschiebung dieses Gleichgewichts durch Überexpression einer der Komponenten des Komplexes zeigte eine Neuverteilung von eu- und heterochromatischen Regionen (Buck und Shore, 1995; Reuter und Spierer, 1992; Eissenberg et al., 1992). Dieser Dosisseffekt kann die Chromatinstruktur an vorbestimmten Loci destabilisie-

ren, was letztlich zu einem Übergang vom normalen zum transformierten Zustand führt.

Trotz der Charakterisierung von HRX/ALL-1 und *bmi* als Protoonkogene, die die Chromatinstruktur verändern können, ist das Wissen über Säugetiergenprodukte, welche mit Chromatin wechselwirken, noch sehr beschränkt. Im Gegensatz dazu wurden durch genetische Analysen von PEV in *Drosophila* ca. 120 Allele für Chromatinregulatoren beschrieben (Reuter und Spierer, 1992). Kürzlich wurde eine carboxyterminale Region mit Ähnlichkeit in der Sequenz identifiziert, die einem positiven (*trx*, *trx*-Gruppe) und eine negativen (*E(z)*, *Pc*-Gruppe) *Drosophila*-Chromatinregulator (Jones und Gelbart, 1993) gemeinsam ist. Darüberhinaus ist dieser Carboxyterminus auch in Su(var)3-9, einem dominanten Suppressor der Chromatinverteilung in *Drosophila*, konserviert (Tschiersch et al., 1994).

Bei der vorliegenden Erfindung wurde von der Überlegung ausgegangen, daß diese als "SET" bezeichnete Proteindomäne (Tschiersch et al., 1994) wegen ihrer evolutionären Konservierung und dem Vorhandensein in antagonistischen Genprodukten eine neue Genfamilie entwicklungsgeschichtlich wichtiger Säugetier-Chromatinregulatoren definiert. Darüberhinaus trägt die Charakterisierung weiterer Mitglieder der Gruppe von SET-Domäne-Genen, neben HRX/ALL-1, zur Aufklärung der Mechanismen bei, die dafür verantwortlich sind, daß strukturelle Veränderungen im Chromatin zur malignen Transformation führen können.

Der vorliegenden Erfindung lag somit die Aufgabe zugrunde, humane Chromatinregulatorgene zu identifizieren, ihre Funktion aufzuklären und sie für die Diagnostik und Therapie einzusetzen.

Zur Lösung dieser Aufgabe wurde zunächst die Sequenzinformation der SET-Domäne benutzt, um aus humanen cDNA-Banken die zu den SET-Domäne-Genen von *Drosophila* homologen humanen cDNAs zu erhalten. Es wurden zwei cDNAs erhalten, die Humanhomologe von *E(z)* bzw. Su(var)3-9 darstellen; die entsprechenden humanen Gene wurden als HEZ-2 und H3-9 bezeichnet (vgl. unten); außerdem wurde eine variante Form von HEZ-2 identifiziert, die als HEZ-1 bezeichnet wurde.

Die vorliegende Erfindung betrifft somit DNA-Moleküle, enthaltend eine für ein Chromatinregulator-Protein, das eine SET-Domäne aufweist, kodierende Sequenz oder eine Teilsequenz davon, dadurch gekennzeichnet, daß sie die in Fig. 6 oder in Fig. 7 dargestellte Nukleotidsequenz aufweisen. Die erfindungsgemäßen DNA-Moleküle werden im folgenden auch als "erfindungsgemäße Gene" bezeichnet.

Die erfindungsgemäßen Gene tragen die Bezeichnung HEZ-2 und H3-9.

Die Erfindung betrifft in einem weiteren Aspekt die von diesen Genen abgeleiteten cDNAs, einschließlich ihrer degenerierten Varianten; Mutanten, die für funktionelle Chromatinregulatoren kodieren sowie Varianten, die auf eine Genduplikation zurückzuführen sind; ein Beispiel dafür ist HEZ-1, dessen Sequenz im Vergleich mit HEZ-2 in Fig. 8 dargestellt ist.

Zur Lösung der im Rahmen der vorliegenden Erfindung gestellten Aufgabe wurde im einzelnen wie folgt vorgegangen: Ausgehend von der Sequenzinformation von der konservierten SET-Domäne, wurde unter reduzierter Stringenz eine humane B-Zell-spezifische cDNA-Bibliothek mit einer gemischten *Drosophila*-DNA-Sonde, die für die SET-Domänen von *E(z)* und Su(var)3-9 kodiert, gescreent. Aus 500.000 Plaques

wurden 40 primäre Phagen ausgewählt. Nach zwei weiteren Runden Screening zeigte sich, daß 31 Phagen für authentische E(z)-Sequenzen kodieren und daß fünf Phagen E(z)-Varianten darstellen. Im Gegensatz dazu hybridisierten nur zwei Phagen mit der Sonde, die die SET-Domäne von Su(var)3-9 allein enthielt. Die Phageninserts wurden mittels Polymerasekettenreaktion (PCR) amplifiziert und mittels Restriktionskartierung und Teilsequenzierung analysiert. Repräsentative cDNA-Inserts wurden

subkloniert und über ihre ganze Länge sequenziert. Die 5'-Enden wurden isoliert, indem positive Phagen noch einmal mit 5'-DNA-Sonden gescreent wurden, worauf nach Subklonierung komplette cDNAs erhalten wurden.

Die komplette, für das humane Homologe von E(z) kodierende cDNA wurde als HEZ-2 und die DNA, die für das humane Homologe von Su(var)3-9 kodiert, wurde als H3-9 bezeichnet. Insgesamt beträgt die Identität der Aminosäuren zwischen *Drosophila* und den humanen Proteinen 61% für HEZ-2 und 43% für H3-9, wobei die C-terminale SET-Domäne sehr hoch konserviert ist (88% für HEZ-2 und 53% für H3-9). Der Sequenzvergleich zeigte weitere deutliche Homologieregionen, z. B. eine Cystein-reiche Domäne in HEZ-2 und eine Chromo-Box in H3-9. (In *polycomb* wurde gezeigt, daß die Chromo-Box die für die Wechselwirkung zwischen DNA und Chromatin wesentliche Domäne ist; Messmer et al., 1992). Im Gegensatz dazu fehlen die 207 Aminosäuren, die das aminoterminal GTP-Bindungsmotiv des *Drosophila*-Proteins enthalten, im putativen humanen Homologen H3-9.

Der Vergleich der Aminosäure-Sequenzen zwischen den *Drosophila*- und den humanen Genen ist in den Fig. 1 und 2 dargestellt:

Fig. 1 zeigt den Vergleich der Aminosäure-Sequenz zwischen HEZ-2 und von *Drosophila*-enhancer von zeste (E(z)). Identische Aminosäuren der konservierten carboxyterminalen SET-Domäne (schattierte Box) und der Cys-reichen Region (Cys-Reste sind hervorgehoben) sind gezeigt. Die mutmaßlichen Kernlokalisierungssignale sind unterstrichen.

Fig. 2 zeigt den Vergleich der Aminosäure-Sequenz zwischen dem humanen Homologen H3-9 und *Drosophila*-Su(var)3-9. Identische Aminosäuren der konservierten carboxyterminalen SET-Domäne (schattierte Box) und der Chromo-Domäne (dunkler schattierte Box) sind dargestellt. Die mutmaßlichen Kernlokalisierungssignale sind unterstrichen. Oben in der Figur ist eine schematische Übersicht der beiden Proteinstrukturen dargestellt, die zeigt, daß im humanen Homologen 207 Aminosäuren am N-Terminus fehlen.

Da translationale Consensussequenzen in Umgebung des Start-ATG der humanen H3-9-cDNA auch an der entsprechenden internen Position in Su(var)3-9 vorhanden sind, dürfte das *Drosophila*-Protein zusätzliche Exons enthalten, die in einem späteren Stadium der Evolution für die Funktion entbehrlich wurden. (Die Richtigkeit dieser Hypothese kann bestätigt werden, indem die humane H3-9-cDNA und komplette oder am 5'-Ende verkürzte cDNAs von Su(var)3-9 in *Drosophila* exprimiert werden. Es wurde eine weitere cDNA der Bezeichnung MG-44 beschrieben (s. unten), der ebenfalls das 5'-Ende fehlt.)

Im Rahmen der vorliegenden Erfindung durchgeführte DNA-Blot-Analysen deuten darauf hin, daß HEZ und H3-9 in der Maus von einzelnen Genen repräsentiert werden, während im Menschen HEZ-Sequenzen durch

zwei getrennte Loci kodiert werden. Der zweite HEZ-Locus (als HEZ-1 bezeichnet) wurde auch durch Charakterisierung einer kleinen Anzahl von cDNA-Varianten, die sich in ihren 3'-flankierenden Sequenzen von der Mehrzahl der aus der humanen cDNA-Bibliothek isolierten HEZ-Klone unterscheiden, bestätigt. Die Unterschiede zwischen HEZ-2 und HEZ-1 im sequenzierten Bereich sind in Fig. 8 dargestellt: Die SET-Domäne von HEZ-1 zeigt gegenüber HEZ-2 Mutationen; außerdem hat HEZ-1 in der SET-Domäne ein Stopcodon, das das Protein um die 47 C-terminalen Aminosäuren verkürzt. In Fig. 8 ist die Nukleotidsequenz der HEZ-2-cDNA von Position 1844 bis 2330 in der jeweils oberen Zeile dargestellt, wobei das authentische Stopcodon unterstrichen ist. Um eine Teilsequenz der cDNA der HEZ-1-Variante der HEZ-2-Sequenz zuzuordnen, wurde das gap-Programm des Wisconsin GCG Netzwerkservice verwendet. Das vorzeitige Stopcodon in HEZ-1 (Position 353) ist unterstrichen. Sequenzen, die für die konservierte SET-Domäne kodieren, sind hervorgehoben. Außerdem ist das 3'-Ende (Position 151 in HEZ-1) des aberranten Transkripts B52 (s. unten) dargestellt. Über die verfügbare Sequenz zeigte sich B52 zu 97% identisch mit HEZ-1 und zu 72% identisch mit HEZ-2. Der Sequenzvergleich von HEZ-1 mit HEZ-2 sowie die Feststellung, daß beim Menschen im Gegensatz zur Maus zwei getrennte HEZ-Loci auftreten, lassen darauf schließen, daß beim Menschen eine Genduplikation aufgetreten sein dürfte.

In einem Vergleich mit cDNA-Sequenzen in der Datenbank GeneBank wurde überraschend festgestellt, daß bestimmte in der Datenbank eingetragene cDNA-Teilsequenzen, die von aberranten Transkripten in Tumorgeweben abgeleitet sind, mutierte Versionen der erfindungsgemäßen cDNAs darstellen:

Einerseits war auf der Suche nach BRCA1, einem Gen, das für Brust- und Eileiterkrebs prädisponiert, eine partielle cDNA-Sequenz mit 271 Nukleotiden der Bezeichnung B52, die für eine mutierte Variante der SET-Domäne kodiert, isoliert und auf dem humanen Chromosom 17q21 kartiert worden (Friedman et al., 1994). Es wurde im Rahmen der vorliegenden Erfindung überraschend festgestellt, daß B52 97% Identität mit der erfindungsgemäßen HEZ-1-cDNA-Variante aufweist (vgl. oben); möglicherweise stellt HEZ-1 ein Pseudogen dar, dessen Reaktivierung bei der deregulierten Proliferation eine Rolle spielt.

Andererseits war eine cDNA (2.800 Nukleotide; MG-44) vom humanen Chromosom Xp11 isoliert worden (Geraghty et al., 1993), einer Region, die für degenerative Störungen der Netzhaut und Synovialsarkome prädisponiert. Es wurde überraschend festgestellt, daß diese cDNA 98% Identität mit der erfindungsgemäßen H3-9-cDNA aufweist. (Der Sequenzvergleich der H3-9-cDNA mit ihrer mutierten Version MG-44 zeigt, daß diese ebenfalls nicht das am 5'-Ende gelegene GTP-Bindungsmotiv des *Drosophila*-Homologen aufweist.)

Die im Rahmen der vorliegenden Erfindung bereitgestellten neuen Gene ermöglichen es somit erstmals, einen Zusammenhang zwischen bestimmten Krebserkrankungen und Mutationen in Chromatinregulatoren herzuleiten; im Falle der MG-44-cDNA konnte, da diese mehrere Punkt- und Frameshift-Mutationen aufweist, welche die Chromo- und SET-Domänen unterbrechen, erst anhand der erfindungsgemäßen H3-9-cDNA ein Zusammenhang zwischen Su(var)3-9 und MG-44 aufgeklärt werden.

Neben den bereits genannten Sequenzen sind in der

Sequenzdatenbank GeneBank als weitere humane Mitglieder der SET-Proteinfamilie das gut dokumentierte humane Homologe von *Drosophila* *trx*, *HRX/ALL-1* (Tkachuk et al., 1992; Gu et al., 1992) eingetragen, ferner ein Gen unbekannter Funktion der Bezeichnung *G9a*, das im humanen Major Histocompatibility Complex (Milner und Campbell, 1993) vorhanden ist, drittens eine nicht publizierte cDNA (*KG-1*), die aus unreifen myeloiden Tumorzellen isoliert wurde (Nomura et al., 1994). Während *G9a* derzeit das einzige humane Gen mit einer SET-Domäne ist, für das bisher keine mutierte Version bekannt ist, trägt *KG-1* eine Insertion von 342 Aminosäuren, welche die SET-Domäne in eine Amino- und eine Carboxyterminale Hälfte spaltet. Ob diese *KG-1*-cDNA das authentische Gen oder eine aberrante Version darstellt, ist unklar. Insgesamt sind vier der fünf derzeit bekannten humanen Mitglieder der SET-Proteinfamilie Änderungen unterworfen, die alle die SET-Domäne mutieren (*HRX/ALL-1*, *HEZ-1/B52*, *MG-44* und *KG-1*). Darüberhinaus wurden in drei Fällen die entsprechenden humanen Genloci in der Nähe von Translokationsbruchpunkten oder instabilen chromosomalen Regionen kartiert (*HRX/ALL-1*, *HEZ-1/B52* und *MG-44*). In Fig. 3 sind die aberranten Transkripte von humanen SET-Domäne-Genen dargestellt. Links in der Figur ist die Lage der fünf derzeit bekannten SET-Domäne-Gene auf dem jeweiligen Chromosom angegeben. U.a. sind die drei Gene (*HRX/ALL-1*, *HEZ-1/B52* und *MG-44*), für die aberrante cDNAs auf Translokationsbruchpunkten oder instabilen Chromatinregionen kartiert wurden, dargestellt. Vier der fünf dargestellten SET-Domäne-Gene weisen Mutationen auf, die alle die Carboxyterminale SET-Domäne unterbrechen, welche in der Figur durch die dunkle Box dargestellt ist. Eine Translokation verbindet die aminoternale Hälfte von *HRX* mit einer nicht korrelierten Gensequenz, die als gepunktete Box mit der Bezeichnung *ENL* dargestellt ist. Ein vorzeitiges Stopcodon verkürzt die SET-Domäne von *HEZ-1/B52*. Punkt- und Frameshift-Mutationen unterbrechen die Chromo- und SET-Domäne in *MG-44*. Eine große Insertion spaltet die SET-Domäne von *KG-1* in zwei Hälften. Aberrante Transkripte sind derzeit für *G9a* nicht bekannt. Das Cystein-reiche Cluster in *B52* ist als gepunktete Box dargestellt; in *HRX/ALL-1* sind die Homologieregion zu Methyltransferasen als schraffierte Box und die A/T-Haken als vertikale Linien dargestellt. Die Namen der jeweiligen authentischen Gene sind rechts in der Figur angegeben.

Die Tatsache, daß ein Säugetiergen der SET-Proteinfamilie, *HRX/ALL-1*, mit translokationsinduzierter Leukämogenese im Zusammenhang gebracht wurde (Tkachuk et al., 1992; Gu et al., 1992), ist ein starker Hinweis dafür, daß Proteine mit SET-Domäne nicht nur wichtige Regulatoren der Entwicklung sind, die chromatinabhängige Veränderungen der Genexpression mitbestimmen, sondern daß sie, nach Mutation, auch die normale Zellproliferation stören.

Da alle bisher beschriebenen Mutationen die Primärstruktur der SET-Domäne unterbrechen, liegt die Vermutung nahe, daß es die SET-Domäne als solche ist, die eine entscheidende Rolle im Übergang vom normalen zum transformierten Zustand spielt. Eine wichtige Funktion für die SET-Domäne kann ferner aufgrund ihrer evolutionären Konservierung in Genprodukten, die von der Hefe bis zum Menschen auftreten, vermutet werden.

Fig. 4 zeigt die evolutionäre Konservierung von SET-Domäne-Proteinen: Unter Verwendung des tfasta-Pro-

gramms des Wisconsin GCG Netzwerkservice wurden Proteine und offene Leserahmen mit Homologie zur SET-Domäne identifiziert. In der Figur ist eine repräsentative Auswahl von Hefe bis zum Menschen gezeigt. Ziffern geben Aminosäuren an. Die Carboxyterminale SET-Domäne ist durch eine schwarze Box dargestellt, Cys-reiche Regionen durch eine dunkel gepunktete Box, das GTP-Bindungsmotiv in *Su(var)3-9* durch eine hellgepunktete Box und die Chromo-Domäne von *Su(var)3-9* und *H3-9* durch eine offene Box mit hellen Punkten. Eine Region mit Homologie zu Methyltransferase (*trx* und *HRX*) ist als schraffierte Box dargestellt. A/T-"Haken" ("A/T Hooks") sind durch vertikale Linien dargestellt. Eine weitere Ser-reiche Region (*S* in *C26E6.10*) und eine Glu-reiche Region (*E* in *G9a*) oder Ankyrin-Repeats (*ANK* in *G9a*) sind ebenfalls hervorgehoben. *YHR119* (GeneBank Accession No. U00059) und *C26E6.10* (GeneBank Accession No. U13875) sind offene Leserahmen von kürzlich in die Datenbank eingetragenen Cosmiden ohne eine funktionelle Charakterisierung. Prozente geben die Gesamtheit der Aminosäure-Identitäten zwischen den humanen und den *Drosophila*-Proteinen an.

Fig. 5 zeigt die Übereinstimmung an Aminosäuren der SET-Domäne. Die SET-Domäne der in Fig. 4 dargestellten Gene wurde unter Verwendung des Pileup-Programms des Wisconsin GCG Netzwerkservice angeordnet. Für den Vergleich der *KG-1*-SET-Domäne wurde vor dem Pileup die große Aminosäure-Insertion, die die SET-Domäne in zwei Hälften spaltet, entfernt. Aminosäure-Positionen, die 8 von 10 Übereinstimmungen aufweisen, sind hervorgehoben.

Aufgrund der im Rahmen der vorliegenden Erfindung festgestellten Kriterien erweist sich eine Beteiligung der Gene, die eine SET-Domäne haben, an der chromatinabhängigen Entstehung der deregulierten Proliferation; diese Gene bzw. die davon abgeleiteten cDNAs sowie Teil- und mutierte Sequenzen können somit bei der Therapie und Diagnose von Erkrankungen, die auf eine derartige Proliferation zurückzuführen sind, eingesetzt werden:

Unterschiede im Transkriptionsniveau von SET-Domäne-RNAs zwischen normalen und transformierten Zellen können als diagnostischer Parameter für Erkrankungen herangezogen werden, in denen die Expression von SET-Domäne-Genen dereguliert ist:

So können Oligonukleotide, kodierend für die SET-Domäne als solche bzw. Teilabschnitte davon, als diagnostischer Marker verwendet werden, um bestimmte Krebsarten, in denen die SET-Domäne mutiert ist, zu diagnostizieren. Für die detaillierte Analyse der Funktion der erfindungsgemäßen cDNAs oder Abschnitten davon im Hinblick auf den diagnostischen Einsatz von SET-Domäne-Gensequenzen wurden im Rahmen der vorliegenden Erfindung die homologen Maus-cDNAs von *HEZ-2* und *H3-9* isoliert. Bei Verwendung der Maus-*HEZ*-cDNA als spezifische DNA-Sonde in "nuclease S1 protection" Analysen zur Untersuchung der *HEZ-2*-Genaktivität während der normalen Mausentwicklung zeigte sich ein eher breites Expressionsprofil, das ähnlich dem von *bmi* (Haupt et al., 1991) ist. Die mit den Maussequenzen durchgeführten Analysen werden mit Humansequenzen ausgeweitet, um die RNA-Mengen zwischen unreifen Vorläuferzellen, Tumorzellen und differenzierten Zellen in verschiedenen humanen Zellkultursystemen zu vergleichen. Um festzustellen, ob sich die SET-Domäne dementsprechend als diagnostischer Tumormarker für spezifische Krebserkrankungen

oder als genereller diagnostischer Parameter eignet, können gängige Methoden zur Bestimmung der RNA-Konzentration, die in einschlägigen Laborhandbüchern (Sambrook, J., Fritsch, E.F. und Maniatis, T., 1989, Cold Spring Harbor Laboratory Press) beschrieben sind, wie Northern Blot, S1-Nuclease-Protection-Analyse oder RNAse-Protection-Analyse, verwendet werden.

Um die Häufigkeit zu untersuchen, mit der die SET-Domäne spezifischen Mutationen unterliegt, können die SET-spezifischen DNA-Sonden zur Analyse von Einzelstrang-Konformationspolymorphismen (single strand conformation polymorphisms; SSCP; Gibbons et al., 1995) eingesetzt werden.

Krebsarten, bei denen SET-spezifische DNA-Sonden als diagnostischer Marker verwendet werden können, sind Brustkrebs (HEZ-2; Friedman et al., 1994), Synovialsarkom (H3-9; Geraghty et al.; 1993) und Leukämien.

Aufgrund der Kenntnis der Nukleotidsequenz der SET-Domäne-Gene können die entsprechenden von der cDNA-Sequenz abgeleiteten Proteine, die ebenfalls Gegenstand der vorliegenden Erfindung sind, in rekombinanter Form hergestellt werden, indem die dafür kodierenden cDNAs in geeignete Vektoren inseriert und in Wirtsorganismen exprimiert werden. Die für die Produktion rekombinanter Proteine verwendeten Techniken sind dem Durchschnittsfachmann geläufig und können einschlägigen Handbüchern (Sambrook, J., Fritsch, E.F. und Maniatis, T., 1989, Cold Spring Harbor Laboratory Press) entnommen werden.

Gegenstand der vorliegenden Erfindung sind somit in einem weiteren Aspekt rekombinante DNA-Moleküle, enthaltend die für HEZ-2, H3-9 oder HEZ-1 kodierende DNA sowie damit funktionell verbundene Expressionskontrollsequenzen, sowie die damit transformierten Wirtsorganismen.

Die erfindungsgemäßen rekombinanten Proteine können eingesetzt werden, um die Wechselwirkung von SET-Domäne-Proteinen mit Chromatin bzw. mit anderen Mitgliedern von Heterochromatinkomplexen zu analysieren; ausgehend von den dabei erhaltenen Erkenntnissen über die Wirkungsweise dieser Komplexe werden die sich im einzelnen ergebenden Möglichkeiten für das gezielte Eingreifen in die daran beteiligten Mechanismen definiert und können für therapeutische Anwendungen genutzt werden.

Untersuchungen, die der weiteren Analyse der Funktion der SET-Domäne dienen, werden z. B. durchgeführt, indem cDNAs, kodierend für humanes HEZ-2 bzw. H3-9 und versehen mit einem Epitop, gegen das Antikörper zur Verfügung stehen, in vitro sowie in Gewebekulturen exprimiert werden. Nach Immunpräzipitation mit den jeweiligen epitopspezifischen Antikörpern kann festgestellt werden, ob HEZ-2 und H3-9 miteinander in vitro wechselwirken können und ob in vivo eine Komplexbildung zwischen HEZ-2 und/oder H3-9 mit weiteren Chromatinregulatoren stattfindet.

Es wurde bereits von anderen Autoren angenommen (DeCamillis et al., 1992; Rastelli et al., 1993; Orlando und Paro, 1993), daß eine Komplexbildung zwischen verschiedenen Mitgliedern von Heterochromatinproteinen wesentlich für deren Funktion ist. Aufgrund der Verfügbarkeit der erfindungsgemäßen SET-Domäne-Gene kann festgestellt werden, ob die SET-Region eine Domäne darstellt, die aufgrund von Wechselwirkungen funktioniert, oder ob sie zum Entstehen multimerer heterochromatischer Komplexe beiträgt. Ebenso kann festgestellt werden, ob die SET-Domäne eine inhibitori-

sche Funktion hat, ähnlich der aminoterminalen BTB-Domäne verschiedener Chromatinregulatoren, einschließlich des GAGA-Faktors (Adams et al., 1992). Insgesamt erlauben die Analysen von Wechselwirkungen mit Epitop-versehenen HEZ-2- und H3-9-Proteinen eine weitere Charakterisierung der Funktion der SET-Domäne. Dadurch werden Möglichkeiten eröffnet, gegen die deregulierte Aktivität vorzugehen, indem z. B. mittels gentherapeutischer Methoden dominant-negative Varianten der SET-Domäne-cDNA-Sequenzen in die Zelle eingeführt werden. Derartige Varianten werden z. B. erhalten, indem zunächst die funktionellen Domänen der SET-Proteine definiert werden, z. B. die für die DNA/Chromatin- oder die für die Protein/Protein-Wechselwirkung verantwortlichen Sequenzabschnitte, und indem dann die um die jeweilige(n) Domäne(n) oder um Abschnitte davon verkürzten DNA-Sequenzen in der betroffenen Zelle zur Expression gebracht werden, um eine durch das intakte funktionelle Protein verursachte Deregulation der Proliferation zu kompetitieren.

Die Verfügbarkeit der erfindungsgemäßen cDNAs erlaubt ferner die Herstellung transgener Tiere, z. B. Mäuse, in denen SET-Domäne-Gene entweder überexprimiert werden können ("gain-of-function") oder in denen diese Gene ausgeschaltet werden können ("loss-of-function"); für letztere Analysen werden die korrespondierenden Tiersequenzen, insbesondere Mausequenzen, der erfindungsgemäßen Gene eingesetzt.

Insbesondere gibt die "gain-of-function"-Analyse, bei der Allele der erfindungsgemäßen Gene in die Maus eingeführt werden, letztlich Aufschluß über die ursächliche Beteiligung von HEZ-2 und H3-9 an der chromatinabhängigen Förderung der Tumorentstehung. Für die "gain-of-function"-Analyse können die kompletten cDNA-Sequenzen von humanem HEZ-2 und H3-9 sowie deren mutierte Versionen, wie HEZ-1/B52 und MG-44, mittels Vektoren, die hohe Expressionsraten ermöglichen, z. B. Plasmiden mit dem humanen β -Actin-Promotor, getrieben sowohl vom Enhancer der schweren Kette von Immunglobulinen (E_μ) als auch von Moloney-Virus-Enhancern (Mo-LTR). Kürzlich wurde gezeigt, daß die E_μ /Mo-LTR-abhängige Überexpression des bmi-Gens, welches gemeinsam mit HEZ-2 zur Pc-Gruppe negativer Chromatinregulatoren zählt, ausreicht in transgenen Mäusen Lymphome zu erzeugen (Alkema et al., 1995).

Indem in "loss-of-function"-Analysen die endogenen Maus-Loci für HEZ-2 und H3-9 durch homologe Rekombination in embryonalen Stammzellen unterbrochen werden, kann bestimmt werden, ob der Verlust der in vivo-Genfunktion zu einer abnormalen Entwicklung der Maus führt.

Aufgrund dieser in vivo-Systeme kann die Wirkung von HEZ-2 und H3-9 bestätigt werden; diese Systeme dienen außerdem als Grundlage für Tiermodelle im Hinblick auf eine humane Gentherapie.

Der gentherapeutische Einsatz der erfindungsgemäßen DNA-Sequenzen oder davon abgeleiteter Sequenzen (z. B. komplementärer Antisenseoligonukleotide) erfolgt — je nachdem, ob die zu behandelnde Erkrankung auf eine Deregulation von Chromatin infolge des Fehlens der funktionellen Gensequenz, oder aber infolge einer Überexpression der entsprechenden Gene oder durch Reaktivierung eines Pseudogens zurückzuführen ist — durch Einführung der funktionellen Gensequenz, durch Inhibierung der Genexpression, z. B. mit Hilfe von Antisenseoligonukleotiden, oder durch Einführung einer für eine dominant-negative Mutante kodierenden

Sequenz. Die Einführung der jeweiligen DNA-Sequenzen in die Zelle kann mit Hilfe von Standardmethoden für die Transfektion höherer eukaryotischer Zellen erfolgen, zu denen der Gentransfer mittels viraler Vektoren (Retrovirus, Adenovirus, Adeno-assoziiertes Virus) oder mittels nicht-viralen Systemen auf Basis der Rezeptor-vermittelten Endozytose zählen; Übersichten über gebräuchliche Methoden werden z. B. von Mitani und Caskey, 1993; Jolly, 1994; Vile und Russel, 1994; Tepper und Mule, 1994; Zatloukal et al., 1993, WO 93/07283 gegeben.

Für eine Inhibierung der Expression der erfindungsgemäßen Gene kommen auch niedermolekulare Substanzen in Betracht, die in die Transkriptionsmaschine eingreifen; nach Analyse der 5'-regulatorischen Region der Gene kann nach Substanzen gescreent werden, die die Wechselwirkung der jeweiligen Transkriptionsfaktoren mit dieser Region ganz oder teilweise blockieren, z. B. mit Hilfe der in der WO 92/13092 beschriebenen Methode.

Eine Inhibierung der deregulierten Proliferation kann auch am Genprodukt ansetzen, indem die entsprechenden Antikörper gegen das HEZ-2- oder H3-9-Protein, vorzugsweise humane oder humanisierte Antikörper, therapeutisch eingesetzt werden. Die Herstellung solcher Antikörper erfolgt nach bekannten Methoden, wie sie z. B. von Malavsi und Albertini, 1992, oder von Rhein, 1993, beschrieben wurden.

Antikörper gegen HEZ-2 oder H3-9, die therapeutisch oder diagnostisch verwendet werden können, sind ebenfalls Gegenstand der vorliegenden Erfindung.

Figurenübersicht

Fig. 1 Aminosäure-Sequenzvergleich zwischen HEZ-2 und E(z),

Fig. 2 Aminosäure-Sequenzvergleich zwischen H3-9 und Su(var)3-9,

Fig. 3 Aberrante Transkripte von humanen SET-Domäne-Genen,

Fig. 4 Evolutionäre Konservierung von SET-Domäne-Proteinen,

Fig. 5 Aminosäure-Übereinstimmung in der SET-Domäne,

Fig. 6 DNA-Sequenz von HEZ-2,

Fig. 7 DNA-Sequenz von H3-9,

Fig. 8 Sequenzvergleich zwischen den cDNAs von humanem HEZ-2 und REZ-1.

Beispiel

a) Herstellung einer cDNA-Bibliothek

Es wurde eine humane B-Zell-spezifische cDNA-Bibliothek, wie von Bardwell und Treisman, 1994, beschrieben, hergestellt, indem Poly(A)⁺-RNA aus humanen BJA-B-Zellen isoliert, mittels Poly(dT)₁₅-Priming revers transkribiert und in doppelsträngige cDNA konvertiert wurde. Nach Zugabe eines EcoRI-Adapters der Sequenz 5' AATCTCTGAGCTCGTCGACA wurde die cDNA in die EcoRI-Stelle des Bakteriophagen λgt10 ligiert. Die Propagierung und Amplifizierung der Bibliothek erfolgte in E.coli C600.

b) Herstellung von DNA-Sonden

Drosophila-DNA-Sonden, kodierend für die konservierten SET-Domänen von E(z) und Su(var)3-9, wur-

den auf der Grundlage der publizierten Drosophila-Sequenzen (Jones und Gelbart, 1993; Tschiersch et al., 1994) mittels Polymerasekettenreaktion (PCR) hergestellt: 1 µg von Drosophila melanogaster-DNA (Clontech) wurde mit den beiden Primern E(z) 1910 (5' ACT-GAATTCGGCTGGGGCATCTTTCTTAAGG) und E(z) 2280 (5' ACTCTAGACAATTTCATTTCACGCTCTATG) einer Pcr-Amplifikation (35 Zyklen zu 30 sek bei 94°C, 30 sek bei 55°C und 30 sek bei 72°C) unterworfen. Die entsprechende SET-Domäne-Sonde für Su(var)3-9 wurde von 10 ng Plasmid-DNA (Tschiersch et al., 1994; Klon M4) mit dem Primerpaar suvar.up (5' ATATAGTACTTCAAGTCCATTCAA-AAGAGG) und suvar.dn (5' CCAGGTACCGTTGGTGCTGTTTAAGACCG) amplifiziert, wobei dieselben Zyklusbedingungen verwendet wurden. Die erhaltenen SET-Domäne-DNA-Fragmente wurden Gel-gereinigt und teilsequenziert, um die Richtigkeit der amplifizierten Sequenzen zu bestätigen.

c) Screenen der cDNA-Bibliothek

5 × 10⁵ Plaque-bildende Einheiten ("plaque forming units", pfu) wurden mit 5 ml Kultur des bakteriellen Wirtsstammes E.coli C600 (bei einer optischen Dichte OD₆₀₀ von 0.5 in 10 mM MgSO₄ suspendiert) bei 37°C 15 min lang inkubiert und dann auf eine große (200 mm × 200 mm) vorgewärmte LB-Schale ausgegossen. Nach Wachstum über Nacht bei 37°C wurden die Phagen auf einer Nylonmembran (GeneScreen) absorbiert. Die Membran wurde, die Seite mit den absorbierten Phagen nach oben schauend, 30 sek lang in Denaturierungslösung (1.5 M NaCl, 0.5 M NaOH) schwimmen gelassen, dann 60 sek in Denaturierungslösung eingetaucht und abschließend 5 min lang in 3 M NaCl, 0.5 M Tris pH 8, neutralisiert. Dann wurde die Membran kurz in 3 × SSC gespült und die Phagen-DNA mittels UV-Vernetzung auf das Nylonfilter fixiert. Das Filter wurde 30 min lang bei 50°C in 30 ml Church-Puffer (1% BSA, 1 mM EDTA und 0.5 M NaHPO₄, pH 7.2) prähybridisiert, anschließend wurden 2 × 10⁶ cpm der radioaktiv markierten DNA-Sondenmischung (E(z)-SET und Su(var)3-9-SET) zugegeben; die DNA-Sonden wurden durch Random-priming unter Verwendung des RediPrime-Kit (Amersham) hergestellt. Die Hybridisierung wurde über Nacht bei 50°C durchgeführt. Nach Entfernung der Hybridisierungslösung wurde das Filter 10 sek in 2 × SSC, 1% SDS bei Raumtemperatur gewaschen, anschließend wurde 10 sek lang bei 50°C gewaschen. Das Filter wurde in Saranwrap gewickelt und unter Verwendung einer Verstärkerfolie der Autoradiographie unterworfen.

Positive Phagenkolonien wurden auf der Originalplatte mittels Zuordnung des Autoradiogramms identifiziert und die entsprechenden Agarstückchen mit dem größeren Ende einer Pasteur-Pipette entfernt. Der Phagen-Pool wurde über Nacht bei 4°C in 1 ml SM-Puffer (5.8 g NaCl, 2 g MgSO₄·H₂O, 50 ml Tris pH 7.5, 5 ml 2%ige Gelatine auf 1 l H₂O), enthaltend einige Tropfen CHCl₃, eluiert. Das Phagenlysat wurde für eine zweite und dritte Screeningrunde wieder ausplattiert um einzelne, gut isolierte positive Plaques (20 bis 100 Plaques pro Platte in der dritten Runde) zu erhalten.

d) Sequenzanalyse

Die cDNA-Inserts von rekombinanten Phagen wurden in den Polylinker von pBluescript KS (Stratagene)

subkloniert und auf einem automatischen Sequenzer (Applied Biosystems) unter Verwendung der Dideoxy-Methode sequenziert. Die komplette Sequenz von wenigstens zwei unabhängigen Isolaten pro erhaltenen Gen wurde mittels "Primer walking" ermittelt. Die Sequenzen wurden mit dem GCG-Software-Paket (University of Wisconsin) analysiert, die Untersuchung auf Homologie wurde mit dem "Blast and fasta" oder "fasta" Netzwerkservice durchgeführt. Die kompletten Sequenzen von HEZ-2 und H3-9 sind in Fig. 6 und 7 dargestellt.

Literatur

- Adams et al., 1992, *Genes & Dev.* 6, 1589—1607. 15
 Alkema et al., 1995, *Nature* 374, 724—727.
 Allshire et al., 1994, *Cell* 76, 157—169.
 Bardwell und Treisman, 1994, *Genes & Dev.* 8, 1644—1677.
 Brunk et al., 1991, *Nature* 353, 351—355. 20
 Buck und Shore, 1995, *Genes & Dev.* 9, 370—384.
 Cleary, 1991, *Cell* 66, 619—622.
 Counter et al., 1992, *Embo J.* 11, 1921—1928.
 DeCamillis et al., 1992, *Genes & Dev.* 6, 223—232.
 Eissenberg et al., 1992, *Genetics* 131, 345—352. 25
 Friedman et al., 1994, *Cancer Research* 54, 6374—6382.
 Garzino et al., 1992, *Embo J.* 11, 4471—4479.
 Geraghty et al., 1993, *Genomics* 16, 440—446.
 Gibbons et al., 1995, *Cell* 80, 837—845. 30
 Gu et al., 1992, *Cell* 71, 701—708.
 Haupt et al., 1991, *Cell* 65, 753—763.
 Jolly, D., 1994, *Cancer Gene Therapy* 1, 51.
 Jones und Gelbart, 1993, *MCB* 13 (10), 6357—6366.
 Kennedy et al., 1995, *Cell* 80, 485—496. 35
 Locke et al., 1988, *Genetics* 120, 181—198.
 Malavsi, F. und Albertini, A., 1992, *TIBTECH* 10, 267—269.
 Messmer et al., 1992, *Genes & Dev.* 6, 1241—1254.
 Milner und Campbell, 1993, *Biochem. J.* 290, 811—818. 40
 Mitani, K. und Caskey, C.T., 1993, *Trends in Biotechnology* 11, 162—166.
 Nomura et al., 1994, Unpublished. GeneBank accession number: D31891.
 Orlando und Paro, 1993, *Cell* 75, 1187—1198. 45
 Pardue, 1991, *Cell* 66, 427—431.
 Rastelli et al., 1993 *Embo J.* 12, 1513—1522.
 Renauld et al., 1993, *Genes & Dev.* 7, 1133—1145.
 Reuter und Spierer, 1992, *BioEssays* 14, 605—612.
 Rhein, R., 1993, *The Journal of NIH Res.* 5, 40—46. 50
 Smouse und Perrimon, 1990, *Dev. Biol.* 139, 169—185.
 Solomon et al., 1991, *Science* 254, 1153—1160.
 Tepper, R.I. und Mule, J.J., 1994, *Human Gene Therapy* 5, 153.
 Tkachuk et al., 1992, *Cell* 71, 691—700. 55
 Tschiersch et al., 1994, *Embo J.* 13 (16), 3822—3831.
 Vile, R. und Russel S., 1994, *Gene Therapy* 1, 88.
 Zatloukal, K., Schmidt, W., Cotten, M., Wagner, E., Stingl, G. und Birnstiel, M.L., 1993, *Gene* 135, 199. 60

Patentansprüche

1. DNA-Moleküle, enthaltend eine für ein Chromatinregulator-Protein, das eine SET-Domäne aufweist, kodierende Sequenz oder eine Teilsequenz davon, **dadurch gekennzeichnet**, daß sie die in Fig. 6, Fig. 7 oder Fig. 8 dargestellte Nukleotidsequenz, kodierend für HEZ-2, H3-9 oder HEZ-1, 65

einschließlich ihrer degenerierten Varianten sowie Mutanten davon, enthalten.

2. DNA-Molekül nach Anspruch 1, dadurch gekennzeichnet, daß es eine cDNA ist.
 3. DNA-Molekül nach Anspruch 1 oder 2, dadurch gekennzeichnet, daß es humanen Ursprungs ist.
 4. cDNA nach Anspruch 3 der Bezeichnung HEZ-2.
 5. cDNA nach Anspruch 3 der Bezeichnung H3-9.
 6. DNA-Molekül nach Anspruch 4, dadurch gekennzeichnet, daß es die für die SET-Domäne von HEZ-2 kodierende Region oder einen Abschnitt davon enthält.
 7. DNA-Molekül nach Anspruch 5, dadurch gekennzeichnet, daß es die für die SET-Domäne H3-9 kodierende Region oder einen Abschnitt davon enthält.
 8. DNA-Molekül nach Anspruch 4, dadurch daß es für eine dominant-negative Mutante von HEZ-2 kodiert.
 9. DNA-Molekül nach Anspruch 4, dadurch daß es für eine dominant-negative Mutante von H3-9 kodiert.
 10. Rekombinantes DNA-Molekül, enthaltend eine in Anspruch 2 definierte cDNA, funktionell verbunden mit Expressionskontrollsequenzen, zur Expression in prokaryotischen oder eukaryotischen Wirtsorganismen.
 11. Prokaryotische oder eukaryotische Wirtsorganismen, transformiert mit rekombinanter DNA nach Anspruch 10.
 12. Rekombinantes humanes Chromatinregulator-Protein HEZ-2 oder ein Abschnitt davon, erhältlich durch Expression einer in Anspruch 4 definierten cDNA.
 13. Rekombinantes humanes Chromatinregulator-Protein H3-9 oder ein Abschnitt davon, erhältlich durch Expression einer in Anspruch 5 definierten cDNA.
 14. Antikörper gegen HEZ-2.
 15. Antikörper gegen H3-9.
 16. Antisense(desoxy)ribonukleotide mit Komplementarität zu einer Teilsequenz einer in Anspruch 1 definierten DNA.
 17. DNA-Molekül, kodierend für die SET-Domäne eines Chromatinregulator-Gens, bzw. für einen Abschnitt davon, zur Behandlung und Diagnose von Erkrankungen des Menschen, die auf eine Deregulation von Chromatinregulator-Genen, die eine SET-Domäne aufweisen, zurückzuführen sind.
 18. DNA-Molekül nach Anspruch 6 oder 7 zur Behandlung und Diagnose von Erkrankungen des Menschen, die auf eine Deregulation von Chromatinregulator-Genen, die eine SET-Domäne aufweisen, zurückzuführen sind.
 19. Antikörper nach Anspruch 14 oder 15 zur Behandlung und Diagnose von Erkrankungen des Menschen, die auf eine Deregulation von Chromatinregulator-Genen, die eine SET-Domäne aufweisen, zurückzuführen sind.
 20. Transgene Maus, enthaltend ein Transgen für die Expression eines Chromatinregulator-Gens, das eine SET-Domäne aufweist, oder einer mutierten Version eines solchen Proteins.
 21. Transgene Maus, erhältlich aus embryonalen Stammzellen, in denen die endogenen Maus-Loci für HEZ-2 und H3-9 durch homologe Rekombination unterbrochen wurden.
 22. Verfahren zum Identifizieren von Säugetier-

Chromatinregulator-Genen, die eine SET-Domäne aufweisen, oder mutierten Versionen davon, dadurch gekennzeichnet, daß man Säugetier-cDNA- oder genomische DNA-Bibliotheken unter Bedingungen niedriger Stringenz mit einem DNA-Molekül, kodierend für die SET-Domäne oder einen Abschnitt davon, hybridisiert. 5

Hierzu 10 Seite(n) Zeichnungen

10

15

20

25

30

35

40

45

50

55

60

65

Fig. 1

```

HEZ-2 — 1 MGQTGGKSEKGPVCRWRKVKSEYMLRQLKRRRADEVKSMFSSNRQKIL 50
      . . . . . | . | : : | | | | : : | | | : | | : . : | : .
E(z) — 1 . . . . . MNSTKVPPPEWKRVRKSEYIKIRQQKRYKRADEIKEAWIRNWDENH 45

51 ERTEILNCEWQKORRIQPVHILTSVSSLRGTRECSVTSDDLDFP..TQVIPL 98
   . . . . | | . | : . | . . . : . . . . | | : | . | : | :
46 HNVQDLYCESKVVQAKPYD....PPHVDCVKRAEVTSYNGIPSGPQKVPI 91

99 KTLNAVASVPIMYSWSPLQQNFVVEDETVLHNIPYMGDEVLDQDGTFFIEE 148
   . : | | . . : | . | . | | | | | | | | | | | | | | | | | | |
92 CVINAVTPIPTMYTWAPTQQNFVVEDETVLHNIPYMGDEVLDKDGKFIEE 141

149 LIKNYDGVKHGDRECGFINDEIFVELVNAL..... 178
   | | | | | | | | : : : : | : | | | | : |
142 LIKNYDGVKHGDKDPSPFMDDAIFVELVHALMRSYSKELEEAAPSTSTAIAK 191

179 .....GQYNDDDDDDDDGDDPE.....EREKQKQDLED.....H 206
   | : | : | : . | . | | : . | . | | .
192 TEPLAKSKQGEGDGVVDVADCESPMKLEKTESKGDLTDVEKKETEETEPVE 241

207 RDDKESRFPK.....FPSDKIFEAISSMFPDKGTAEELKEKYKELTE 249
   : | : : : . | . . | : | | . | | | | : | | | | | |
242 TEDADVKAEEVVKDLPFPAPIIFQAISANFPDKGTAQELKEKYIELTE 291

250 QQLPGALPPECTPNIDGPNKSVQREQSLHSFHTLPRFKYDPLHFP 299
   : | : . | | | | | . | . | | . . | | | | | | | | | | | | :
292 HQDPER.PPECTPNIDGKAESVSRRRTYHSFHTLPRFKYDPLHRL 340

300 ..HATPNTYKRKNTETALDNKPGPOCYCHLEGAKEFAAALTAERIKTPP 347
   | | . | | : : . | . . | : | | : | : | | | | | |
341 QGHAGPNLQKRRYPELKPFAPESNSCYMLIDGMKEKLAA....DSKTPP 386

348 KRPGGRRGRPLPNSSRPSTPTINVLESKOTDSCREAGTETGGENNDKEE 397
   : . . . . | : . . . . . | . | : .
387 .....IDSCNEASSEDSDNSNSQFSNKDFNH 412

398 EEKDET.SSSSEANSRCQTPIKMKPNIEPPENVEWSGAEASMFVRLIGT 446
   | : | : . . | . | . . . : . | . . . | | : | : | | :
413 ENSKDNGLTVNSAAVAEINSIMAGMMNITSTQCV.WTGADQALYRVLHKV 461

447 YYDNFAIARLIGTKTRQVYEFVRKESSIIAPAPAEVDTPPRKKKKKH 496
   | . : | | | : : | | | | | | | | . . . : : | | | | : :
462 YLKNYCAIAHNMLTKTRQVYEFQAKEDAESFSEDLRQDFTPPRKKKKKQ 511

497 RLWAAHAKIQLKKDGSSNHVYNYQPDHPROPDSSCPVIAQNFCEKF 546
   | | . | | | | | | | | | | | | | | | | . | . | | | |
512 RLWSLHAKIQLKKDSSSNHVYNYTPDHPGHEPLMCPDIQTNFCEKF 561

547 GCSSECNRFPGCKAKCNTKCPYLAVRECPDILCTGAADHWDS 596
   | | | : | | | | | | | | | | | | | | | | | | | |
562 GCSSECNRFPGCKAKCNTKCPYLAVRECPDILCTGADQFGL 610

597 KNVSCNCSIQRGSKKHL LLPSDVAGWGIFIKDPVOKNEFISEYCGEII 646
   . . . | | | | : | : | | | | | | | | | | | | | |
611 TKITCNVTQRGLHKHLLMAPSDIAGWGIFLKEGAQKNEFISEYCGEII 660

647 SQDEADRRGKVYDKYMCNFLNLNDFVVDATRKGNKIRFANHSVNPNCY 696
   | : | | | | | | | | | | | | | | | | | | | |
661 SQDEADRRGKVYDKYMCNFLNLNDFVVDATRKGNKIRFANHVINPNCY 710

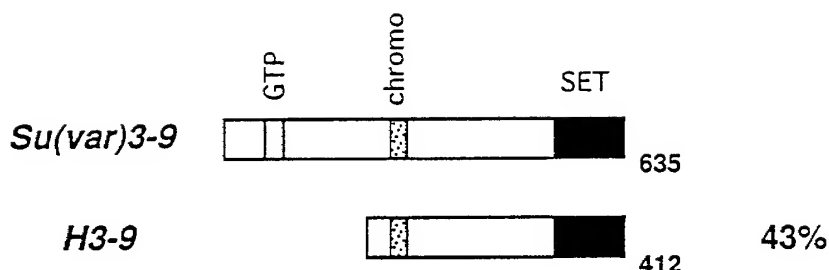
697 AKVMNVNGEHRIGIFAKRAIQTGEELFFDYTRYQADALKYVGIEREMIEIP 746
   | | | | | | | | | | | | | | | | | | | | | |
711 AKVMNVGTSHRIGIFAKRAIQPEELFFDYRYGPTSQLKFVGIEREMIEIV 762

```

C-reich
75%

SET
88%

Fig. 2



H3-9 — 1 MAENLKGC SVCKSSWNQLQDLCLRLAKLSCPALGISKRNLVDFEVEYLCD 50

Su(var)3-9 — 207MGVIAKREPKG.....EYVERIEC 226

51 YKKIRECEYLLVKNRGYPDSSTHPEPCNLK..... 81

227 VEMDQYQPVFEFKWLGYSNTHESLANVADCAEMEXFVERHQQLYETY 276

82 CVRLKQFHKDLERELLRRHRSKTPHLDPLSLANYLVQAKQRALRR 130

277 IAKITTELEKQLEALPLMENITVAEVDAYEPLNLQIDLILLAQYRAAGSR 326

131 WEQE.....LNAKRSHLGR.....IT 146

327 SQREPQKIGERALKSMQIKRAQFVRRKQLADLALFEKRMNHVEKPSPPIR 376

147 VENEVDLDGPPRAFYINEYRVGEGITLNQVA.VGDDCLWAPTGG 195

377 VENNIDLDTIDSNFMYIHDNIIGKDVPKPEAGIVGTTEDTEECTASTK 426

196 PGASL..HKFAYNDQG.QVRLRAGLP IYENSRGCEYDGNRVVQKGIR 242

427 GQAFAGELFAYERSTRRLRLPGSAIYENSRGCEYDGNRVVQKGIR 476

243 YDLCIFRTDDGRGAGVRLTKIRKNSFVMEYVGEIITSEEAERGGIYDR 292

477 VPLVLFKTANGSGNGVRAATAALRKGEFVCEYIEEYITSDANERGGKAYDD 526

293 QGATYLFDLDY...VEDVYTVDAAYYGNISHFVNHSCDPNLQVNVTFIDN 339

527 NGRTYLFDLDYNTAQDSEXTIDAANYGNISHFVNHSCDPNLAVFPCWIEH 576

340 LDERLPRIAFFATRTIRAGEELTFDYNMQVDPVDMESTRMDSNFGLAGLP 389

577 LNVALPHLVFFTLRPIKAGEELSEDY.IRADNEDVPYENLSTA..... 619

390 GSPKFRVRIECKCGTESCKRYLF 412

619VRVECRGGRDNCRKVLF 635

CHROMO
39%

SET
51%

Fig. 3

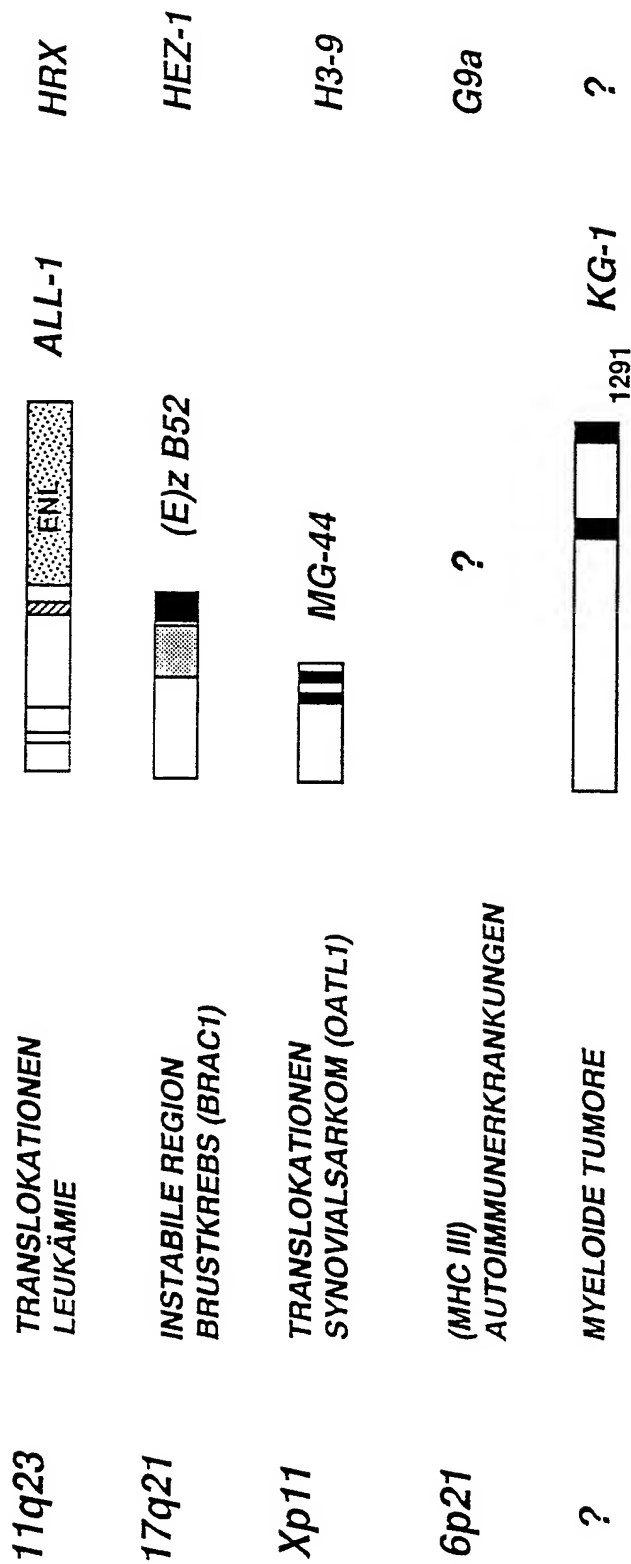


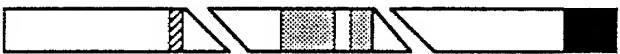


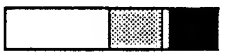
Fig. 4

S. CEREVISIAE YHR119  1080

C. ELEGANS C26E6.10  739

DROSOPHILA M.

trx  3751

E(z)  760

Su(var)3-9  635

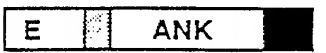
S E T

SÄUGETIERE

HMG I/Y
HRX  3969 46%

HEZ-2  746 61%

H3-9  412 43%

G9a  1001
HMG-1

?

Fig. 5

E(z)	SDIAGWGIFL	KEGAQKNEFI	SEYCGEIIISQ	DEADRRGKVY	DK..YMCSFL	50
HEZ	SDVAGWGIFI	KDPVQKNEFI	SEYCGEIIISQ	DEADRRGKVY	DK..YMCSFL	
HRX	SPIHGRGLFC	KRNIDAGEMV	IEYAGNVIRS	IQTDKREKYY	DSKGIG.CYM	
trx	SHIHGRGLYC	TKDIEAGEMV	IEYAGELIRS	TLTDKRERYY	DSRGIG.CYM	
C26	SRIHGWGLYA	MESIAPDEMI	VEYIGQTIRS	LVAEREKAY	ERRGIGSSYL	
YHR	SAIHNWGLYA	LDSIAAKEMI	IEYVGERIRQ	PVAEMREKRY	LKNGIGSSYL	
su3-9	ANGSGWGVRA	ATALRKGEFV	CEYIEEIIITS	DEANERKAY	DDNG..RTYL	
H3-9	DDGRGWGVRT	LEKIRKNSFV	MEYVGEIITS	EEAERRGQLY	DRQG..ATYL	
G9a	TAKMGWGVRA	LQTIPQGTFI	CEYVGELISD	AED...V...	.RED..DSYL	
KG-1	TQNKGWGIRC	LDDIAKGSFV	CIYAGKIITD	DFADKEGL.	.EMG..DEYF	

E(z)	FNLN.....	NDFVVDATRK	GNKIRFANHS	INPNCYAKVM	MVTGDH....	100
HEZ	FNLN.....	NDFVVDATRK	GNKIRFANHS	VNPNCYAKVM	MVNGDH....	
HRX	FRID.....	DSEVV DATMH	GNRAREINHS	CEPN CYSRVI	NIDGQK....	
trx	FKID.....	DNLVVDATMR	GNAAREINHC	CEPN CYSKVV	DILGHK....	
C26	FRID.....	LHHVIDATKR	GNFAREINHS	QPN CYAKVL	TIEGEK....	
YHR	ERV D.....	ENTVIDATKK	GGIAREINHC	CDPNCTAKII	KVGRR....	
su3-9	FDLDYNTAQD	SEYTIDAANY	GNISHEINHS	CDPNLAVFPC	WIEHLNVALP	
H3-9	FDLDY...VE	DVYTVDAAYY	GNISHEVNHS	CDPNLQVYNV	FIDNLDERLP	
G9a	FDLDNK..DG	EVYCIDARYY	GNISREINHL	CDPNIIIPVRV	FMLHQDLRFP	
KG-1	ANLDHI..ES	VEYIIDAKLE	GNLGRYLNHS	CSPNLFVQNV	FVTHDLRFP	

E(z)	RIGIFAKRAI	QPGEELFFDY	..RYGPTQL	K....FVGI	EREMEIV*	150
HEZ	RIGIFAKRAI	QTGEELFFDY	..RYSQADAL	K....YVGI	EREMEIP*	
HRX	HIVIFAMRKE	YRGEELTYDY	..KFPIE.DA	SNKLPCNCGA	KKCRKFLN*	
trx	HLIIFAVRRE	VQGEELTYDY	..KFPFE.D.	.EKIPCSCGS	KRCRKYLN*	
C26	RIVIYSRIII	KKGEELTYDY	..KFPIE...	DDKIDCLCGA	KTCRGYLN*	
YHR	RIVIYALRDE	AASEELTYDY	..KFEREKDD	EERLPCLCGA	PNCKGFLN*	
su3-9	HLVFFTLRPI	KAGEELSFDY	.IRADNEDVP	YENLSTA...	
H3-9	RIAFFATRTI	RAGEELTFDY	NMQVDPVDME	STRMDSNFGI	AGLPGSPKKR	
G9a	RIAFFESSRDI	RTGEELGFDY	GDRFW..DIK	SKYFTQCQGS	EKCKHSAEAI	
KG-1	WVAFFASKRI	RAGTELTWDY	NYEVG..SVE	GKELLCCCGA	IECR.....	

E(z)
HEZ

HRX
trx
26
YHR

su3-9	VRVECRCGRD	NCRKVLF*
H3-9	VRIECKCGTE	SCRKYLF*
G9a	ALEQSRLARL	DPHPPELLPEL GSLPPVNT*
KG-1GRLL*	

Fig. 6/1

HEZ-2 Length: 2607bp (CODING: 90-2330)

```

1  AGGCAGTGGG GCGGCGGCGG CGGCGGCGGC GCGGCGCGGG GCGGACGCGC
51  GGGAAACAACG CGAGTCGGCG CGCGGGACGA AGAATAATCAM TGGGCCAGAC
101 TGGGAAGAAA TCTGAGAAGG GACCAGTTTG TTGGCGGAAG CGTGTAATAAT
151 CAGAGTACAT GCGACTGAGA CAGCTCAAGA GGTCAGACG AGCTGATGAA
201 GTAAAGAGTA TGTTTAGTTC CAATCGTCAG AAAATTTTGG AAAGAACGGA
251 AATCTTAAAC CAAGAATGGA AACAGCGAAG GATACAGCCT GTGCACATCC
301 TGACTTCTGT GAGCTCATTG CGCGGGACTA GGGAGTGTTT GGTGACCAGT
351 GACTTGGATT TTCCAACACA AGTCATCCCA TTAAAGACTC TGAATGCAGT
401 TGCTTCAGTA CCCATAATGT ATTCTTGGTC TCCCCTACAG CAGAATTTTA
451 TGGTGGAAGA TGAAACTGTT TTACATAACA TTCCTTATAT GGGAGATGAA
501 GTTTTAGATC AGGATGGTAC TTTCATTGAA GAACTAATAA AAAATTATGA
551 TGGGAAAGTA CACGGGGATA GAGAATGTGG GTTTATAAAT GATGAAATTT
601 TTGTGGAGTT GGTGAATGCC CTTGGTCAAT ATAATGATGA TGACGATGAT
651 GATGATGGAG ACGATCCTGA AGAAAGAGAA GAAAAGCAGA AAGATCTGGA
701 GGATCACCGA GATGATAAAG AAAGCCGCCC ACCTCGGATA TTTCTTCTG
751 ATAAAAATTTT TGAAGCCATT TCCTCAATGT TTCCAGATAA GGGCACAGCA
801 GAAGAACTAA AGGAAAAATA TAAAGAACTC ACCGAACAGC AGCTCCCAGG
851 CGCACTTCCT CCTGAATGTA CCCCCAACAT AGATGGACCA AATGCTAAAT
901 CTGTTTCAGAG AGAGCAAAGC TTCACTCCT TTCATACGCT TTTCTGTAGG
951 CGATGTTTTA AATATGACTG CTTCTACAT CCTTTTCAAG CAACACCCAA
1001 CACTTATAAG CGGAAGAACA CAGAAACAGC TCTAGACAAC AAACCTTGTG
1051 GACCACAGTG TTACCAGCAT TTGGAGGGAG CAAAGGAGTT TGCTGCTGCT
1101 CTCACCGCTG AGCGGATAAA GACCCACCA AAACGTCCAG GAGGCCGCAG
1151 AAGAGGACGG CTTCCCAATA ACAGTAGCAG GCCCAGCACC CCCACCATTA
1201 ATGTGCTGGA ATCAAAGGAT ACAGACAGTG ATAGGGAASC AGGGACTGAA
1251 ACGGGGGGAG AGAACAATGA TAAAGAAGAA GAAGAGAAGA AAGATGAAAC
1301 TTCGAGCTCC TCTGAAGCAA ATTCTCGGTG TCAAACACCA ATAAAGATGA
1351 AGCCAAATAT TGAACCTCCT GAGAATGTGG AGTGGAGTGG TGCTGATGOC

```


Fig. 6/2

1401 TCAATGTTTA GAGTCCTCAT TGGCACTTAC TATGACAATT TCTGTGCCAT
1451 TGCTAGGTTA ATTGGGACCA AAACATGTAG ACAGGTGTAT GAGTTTAGAG
1501 TCAAAGAATC TAGCATCATA GCTCCAGCTC CCGCTGAGGA TGTGGATACT
1551 CCTCCAAGGA AAAAGAAGAG GAAACACCGG TTGTGGGCTG CACACTGCAG
1601 AAAGATACAG CTGAAAAAGG ACGGCTCCTC TAACCATGTT TACAACTATC
1651 AACCCTGTGA TCATCCACGG CAGCCTTG TG ACAGTTCGTG CCCTTG TG TG
1701 ATAGCACAAA ATTTTGTGA AAAGTTTGT CAATGTAGTT CAGAGTGTCA
1751 AAACCGCTTT CCGGGATGCC GCTGCAAAGC ACAGTGCAAC ACCAAGCAGT
1801 GCCCGTGCTA CCTGGCTGTC CGAGAGTGTG ACCCTGACCT CTGTCTTACT
1851 TGTGGAGCCG CTGACCATTG GGACAGTAAA AATGTGTCCT GCAAGAACTG
1901 CAGTATTCAG CGGGGCTCCA AAAAGCATCT ATTGCTGGCA CCATCTGACG
1951 TGGCAGGCTG GGGGATTTTT ATCAAAGATC CTGTGCAGAA AAATGAATTC
2001 ATCTCAGAAT ACTGTGGAGA GATTATTTCT CAAGATGAAG CTGACAGAAG
2051 AGGGAAAGTG TATGATAAAT ACATGTGCAG CTTTCTGTTC AACTTGAACA
2101 ATGATTTTGT GGTGGATGCA ACCCGCAAGG GTAACAAAAT TCGTTTTGCA
2151 AATCATTCGG TAAATCCAAA CTGCTATGCA AAAGTTATGA TGGTTAACGG
2201 TGATCACAGG ATAGGTATTT TTGCCAAGAG AGCCATCCAG ACTGGCGAAG
2251 AGCTGTTTTT TGATTACAGA TACAGCCAGG CTGATGCCCT GAAGTATGTC
2301 GGCATCGAAA GAGAAATGGA AATCCCTTGA CATCTGCTAC CTCCTCCCCC
2351 TCCTCTGAAA CAGCTGCCTT AGCTTCAGGA ACCTCGAGTA CTGTGGGCAA
2401 TTTAGAAAAA GAACATGCAG TTTGAAATTC TGAATTTGCA AAGTACTGTA
2451 AGAATAATTT ATAGTAATGA GTTTAAAAAT CAACTTTTTA TTGCCCTTCTC
2501 ACCAGCTGCA AAGTGTTTTG TACCAGTGAA TTTTGTGCAAT AATGCAGTAT
2551 GGTACATTTT TCAACTTTGA ATAAAGAATA CTTGAACTTG TCAAAAAAAA
2601 TGTCGAC

Fig. 7/1

H3-9 LENGTH: 2747bp (CODING: 53-1291)

1 GTCGACACTC GCGAGGCCGG CTAGGCCCGA ATGTCGTTAG CCGTGGGGAA
M
51 AGATGGCGGA AAATTTAAAA GGCTGCAGCG TGTGTTGCAA GTCTTCTTGG
101 AATCAGCTGC AGGACCTGTG CCGCCTGGCC AAGCTCTCCT GCCCTGCCCT
151 CGGTATCTCT AAGAGGAACC TCTATGACTT TGAAGTCGAG TACCTGTGCG
201 ATTACAAGAA GATCCGCGAA CAGGAATATT ACCTGGTGAA ATGGCGTGGA
251 TATCCAGACT CAGAGAGCAC CTGGGAGCCA CGGCAGAATC TCAAGTGTGT
301 GCGTATCCTC AAGCAGTTCC ACAAGGACTT AGAAAGGGAG CTGCTCCGGC
351 GGCACCACCG GTCAAAGACC CCCCGGCACC TGGACCCAAG CTTGGCCAAC
401 TACCTGGTGC AGAAGGCCAA GCAGAGGCGG GCGCTCCGTC GCTGGGAGCA
451 GGAGCTCAAT GCCAAGCGCA GCCATCTGGG ACGCATCACT GTAGAGAATG
501 AGGTGGACCT GGACGGCCCT CCGCGGGCCT TCGTGACAT CAATGAGTAC
551 CGTGTGGTG AGGGCATCAC CCTCAACCAG GTGGCTGTGG GCTGCGAGTG
601 CCAGGACTGT CTGTGGGCAC CCACTGGAGG CTGCTGCCCC GGGGCGTCAC
651 TGCACAAGTT TGCCTACAAT GACCAGGGCC AGGTGCGGCT TCGAGCCGGG
701 CTGCCCATCT ACGAGTGCAA CTCCCGCTGC CGCTGCGGCT ATGACTGCCC
751 AAATCGTGTG GTACAGAAGG GTATCCGATA TGACCTCTGC ATCTTCCGGA
801 CGGATGATGG GCGTGGCTGG GCGTCCGCA CCCTGGAGAA GATTCGCAAG
851 AACAGCTTCG TCATGGAGTA CGTGGGAGAG ATCATTACCT CAGAGGAGGC
901 AGAGCGGCGG GGCCAGATCT ACGACCGTCA GGGCGCCACC TACCTCTTTG
951 ACCTGGACTA CGTGGAGGAC GTGTACACCG TGGATGCCGC CTACTATGGC
1001 AACATCTCCC ACTTTGTCAA CCACAGTTGT GACCCCAACC TGCAGGTGTA
1051 CAACGTCTTC ATAGACAACC TTGACGAGCG GCTGCCCCGC ATCGCTTTCT
1101 TTGCCACAAG AACCATCCGG GCAGGCGAGG AGCTCACCTT TGATTACAAC
1151 ATGCAAGTGG ACCCCGTGGA CATGGAGAGC ACCCGCATGG ACTCCAACCT
1201 TGGCCTGGCT GGGCTCCCTG GCTCCCCTAA GAAGCGGGTC CGTATTGAAT
1251 GCAAGTGTGG GACTGAGTCC TGCCGCAAAT ACCTCTTCTA * GCCCTTAGAA
1301 GTCTGAGGCC AGACTGACTG AGGGGGCCTG AAGCTACATG CACCTCCCCC
1351 ACTGCTGCCC TCCTGTGAG AATGACTGCC AGGGCCTCGC CTGCCTCCAC

Fig. 7/2

1401 CTGCCCCAC CTGCTCCTAC CTGCTCTACG TTCAGGGCTG TGGCCGTGGT
1451 GAGGACCGAC TCCAGGAGTC CCCTTTCCCT GTCCCAGCCC CATCTGTGGG
1501 TTGCACTTAC AAACCCCCAC CCACCTTCAG AAATAGTTTT TCAACATCAA
1551 GACTCTCTGT CGTTGGGATT CATGGCCTAT TAAGGAGGTC CAAGGGGTGA
1601 GTCCCAACCC AGCCCCAGAA TATATTTGTT TTTGCACCTG CTTCTGCCTG
1651 GAGATTGAGG GGTCTGCTGC AGGCCTCCTC CCTGCTGCCC CAAAGGTATG
1701 GGGAAGCAAC CCCAGAGCAG GCAGACATCA GAGGCCAGAG TGCCTAGCCC
1751 GACATGAAGC TGGTTCCCCA ACCACAGAAA CTTTGTACTA GTGAAAGAAA
1801 GGGGTCCCTG GCCTACGGGC TGAGGCTGGT TTCTGCTCGT GCTTACAGTG
1851 CTGGGTAGTG TTGGCCCTAA GAGCTGTAGG GTCTCTTCTT CAGGGCTGCA
1901 TATCTGAGAA GTGGATGCCC ACATGCCACT GGAAGGGAAG TGGGTGTCCA
1951 TGGGCCACTG AGCAGTGAGA GGAAGGCAGT GCAGAGCTGG CCAGCCCTGG
2001 AGGTAGGCTG GGACCAAGCT CTGCCTTCAC AGTGCAGTGA AGGTACCTAG
2051 GGCTCTTGGG AGCTCTGCGG TTGCTAGGGG CCCTGACCTG GGGTGTCTATG
2101 ACCGCTGACA CCACTCAGAG CTGGAACCAA GATCTAGATA GTCCGTAGAT
2151 AGCACTTAGG ACAAGAATGT GCATTGATGG GGTGGTGATG AGGTGCCAGG
2201 CACTAGGTAG AGCACCTGGT CCACGTGGAT TGTCTCAGGG AAGCCTTGAA
2251 AACCACGGAG GTGGATGCCA GGAAAGGGCC CATGTGGCAG AAGGCAAAGT
2301 ACAGGCCAAG AATTGGGGGT GGGGGAGATG GCTTCCCCAC TATGGGATGA
2351 CGAGGCGAGA GGGAAGCCCT TGCTGCCTGC CATTCCCAGA CCCAGCCCT
2401 TTGTGCTCAC CCTGGTTCCA CTGGTCTCAA AAGTCACCTG CCTACAAATG
2451 TACAAAAGGC GAAGGTTCTG ATGGCTGCCT TGCTCCTTGC TCCCCACCC
2501 CCTGTGAGGA CTTCTCTAGG AAGTCCTTCC TGACTACCTG TGCCCAGAGT
2551 GCCCCTACAT GAGACTGTAT GCCCTGCTAT CAGATGCCAG ATCTATGTGT
2601 CTGTCTGTGT GTCCATCCCG CCGGCCCCCC AGACTAACCT CCAGGCATGG
2651 ACTGAATCTG GTTCTCCTCT TGTAACCCCC TCAACCCTAT GCAGCCTGGA
2701 GTGGGCATCA ATAAAAAGAA CTGTCGACTG AAAAAAAAAA TGTCGAC

Fig. 8

HEZ-2 — 1844 TCTTACTTGTGGAGCCGCTGACCATTGGGACAGTAAAAATGTGTCCTGCA 1893
 || || |||| ||| | || || ||||| | || ||||| |
 HEZ-1 — 1 ACTCACCTGTGGGGCCTCAGAGCACTGGGACTGCAAGGTGGTTTCCTGTA 50
 1894 AGAACTGCAGTATTCAGCGGGGCTCCAAAAAGCATCTATTGCTGGCACCA 1943
 | ||||| || || ||||| || ||||| || ||||| ||
 51 AAAACTGCAGCATCCAGCGTGGACTTAAGAAGCACCTGCTGCTGGCCCCC 100
 1944 TCTGACGTGGCAGGCTGGGGGATTTTATCAAAGATCCTGTGCAGAAAAA 1993
 ||||| ||||| || ||||| | || || || || || ||||| || ||||| ||
 101 TCTGATGTGGCCGGATGGGGCACCTTCATAAAGGAGTCTGTGCAGAAGRA 150
 B52 — 1994 TGAATTCATCTCAGAATACTGTGGAGAGATTATTTCTCAAGATGAAGCTG 2043
 ||||| || ||||| ||||| ||| | || ||||| ||||| |||||
 151 CGAATTCATTTCTGAATACTGTGGTGAGCTCATCTCTCAGGATGAGGCTG 200
 2044 ACAGAAGAGGGAAAGTGTATGATAAAATACATGTGCAGCTTTCTGTTCAAC 2093
 | || | || || || ||||| ||||| ||||| ||||| ||||| |||||
 201 ATCGACCGCGAAAGGTCTATGACAAATACATGTCCAGCTTCCTCTTCAAC 250
 2094 TTGAACAATGATTTTGTGGTGGATGCAACCCGCAAGGGTAACAAAATTTCG 2143
 || || ||||| ||||| || || || || || ||||| |||||
 251 CTCAATAATGATTTGTAGTGGATGCTACTCGGAAAGGAACAAAATTTCG 300
 2144 TTTTGCAAAATCATTCCGTAATCCAACTGCTATGCRAAAGTT..ATGAT 2191
 ||||| ||||| || ||||| ||||| ||||| ||||| ||||| | |
 301 ATTTGCAAAATCATTCACTGAATCCCACTGTTATGCCAAAGGTGAGTCCC 350
 2192 GGTAAACGGTGATCACAGGATAGGTATTTTGCCAAGAGAGCCATCCAGA 2241
 || || || || || || || || || || || || || || || || || ||
 351 AGTAACCTGGGAGGTGGGGTGGGGGATGGATGCCTCTTTACTGTGATTC 400
 2242 CTGGCGAAGAGCTGTTTTTTGATTACAGATACAGCCAGGCTGATGCCCTG 2291
 | || | || || || || || || || || || || || || || || || ||
 401 CATTCGTGTGTAACATTTTCCTTAGCTGAGCTATCTTTGTCCAAAGAT 450
 2292 AAGTATGTGCGCATCGAAAGAGAAATGGAAATCCCTTGA 2330
 || || | || || || || || || || || || || || || || || || ||
 451 AATCATGATTAATATCTGGTATCATTTTAGGCCCTCTC 489